POLITECNICO DI TORINO
Corso di Laurea Magistrale in Ingegneria Elettrica
Relatore: Prof. Gianfranco Chicco
Candidato: Richard Andree Toledo Barraza

Analisi delle curve di carico: Confronti tra metodi di clustering con dati di ingresso basati sulle curve di durata

INTRODUZIONE

Il sempre maggiore impatto dell'energia elettrica nella vita delle persone ha richiesto l'utilizzo di nuovi strumenti in grado di facilitare l'analisi dei consumi energetici, soprattutto nel caso di utenti residenziali i quali presentano una elevata variabilità nei consumi e sono quindi molto più complicati da analizzare. Uno degli strumenti che rende più semplice lo studio delle curve di carico di questa tipologia di utenti sono gli algoritmi di clustering, che grazie alla loro capacità di raggruppamento, permettono di creare delle classi in cui gli utenti presentano delle similitudini sulle curve di carico. L'approccio che si utilizza per estrarre l'informazione dalla curve di carico non è quello classico, ma è un metodologia introdotta recentemente, chiamata CONDUCT (CONsumption DUration Curves Times Series), la quale utilizza le curve di durata, in cui non è richiesta la sequenza temporale dei consumi degli utenti durante la giornata. Le curve di durata sono utilizzate in una fase pre-clustering. L'obiettivo di questa tesi è identificare per quale algoritmo di clustering la costruzione dei dati con CONDUCT dà i risultati migliori.

ALGORITMI DI CLUSTERING

Gli obiettivi che gli algoritmi di clustering cercano di ottimizzare sono principalmente due:

- Avere elementi all'interno dei cluster il più simili possibile.
- Avere dei cluster abbastanza dissimili tra di loro.

Il modo in cui questi obiettivi vengono portati a termine non sono uguali, ma dipendono dalla tipologia di algoritmo che si sta considerando.

Gli algoritmi di interesse per questa trattazione sono:

- K-means: E' un algoritmo di clustering euristico appartenente alla famiglia dei Centroid-based clustering. I punti di partenza per la formazione dei cluster sono scelti in modo casuale dal dataset considerato, i quali serviranno da centroidi iniziali per determinare la vicinanza con il resto dei punti.
- Clustering gerarchico: Questa tipologia di clustering appartiene alla famiglia dei Connectivity-based clustering. Come indicato dal nome il principio sul quale si basa è la formazione di un albero binario gerarchico in cui alla base si hanno tutti i punti del dataset considerati come dei cluster singoli, i quali sono raggruppati due per volta in base alla vicinanza e al criterio di linkage scelto.
- DBSCAN: Questo algoritmo appartiene al gruppo dei Density-based clustering. Il principio sul quale si basa è l'identificazione di aree ad elevata densità di punti, le quali sono considerate come cluster. Per essere definite tali le aree devono rispettare un certo raggio di ricerca e avere un dato numero di elementi dentro quel raggio.
- Spectral clustering: E' un tipo di algoritmo appartenente alla famiglia dei Graph-based clustering, il quale utilizza un grafo di similarità per rappresentare i punti del dataset in uno spazio dimensionale ridotto. A partire dal grafo di similarità si ricava la matrice laplaciana rappresentativa del grafo, i cui autovettori servono per rappresentare i dati nello spazio ridotto.

FASI PRE-CLUSTERING

Prima di poter eseguire gli algoritmi di clustering è necessario rendere i dati i più affidabili possibili, questo lo si ottiene attraverso le fasi pre-clustering:

- Misura dati.
- Scelta delle condizioni di carico.
- Pulizia dei dati.

Questi passaggi sono in comune sia all'approccio classico sia a CONDUCT, la differenziazione avviene nel modo in cui si vanno a scegliere le curve rappresentative:

- Approccio classico: Per ogni cliente questa viene ricavata attraverso metodi statistici, utilizzando tutte le curve di carico giornaliere nell'arco temporale considerato.

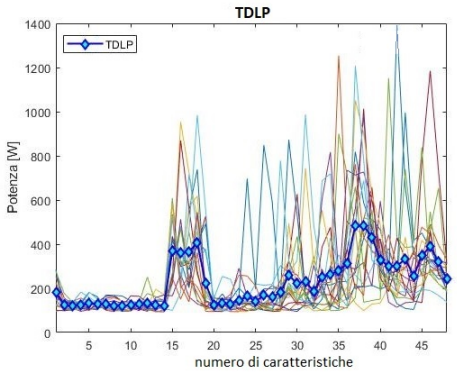

Figura 1: TLDP ricavata dalle curve di carico giornaliere di un mese, mediando i valori nei rispettivi momenti della giornata.

- Metodologia di CONDUCT: I valori appartenenti alla finestra temporale considerata sono tutti disposti in modo crescente per ottenere un unico vettore per utente in cui i valori del consumo seguono una sequenza ordinata, che prende il nome di curva di durata. I punti delle curve di durata da utilizzare per ogni cliente vengono determinati utilizzando la funzione di distribuzione cumulativa (CDF) delle differenze medie, la quale è stata ricavata mediando le differenze tra valori adiacenti in ciascuna curva di durata. Sarà la CDF a indicare i 9 punti (decili) da prendere dalle curve di durata.

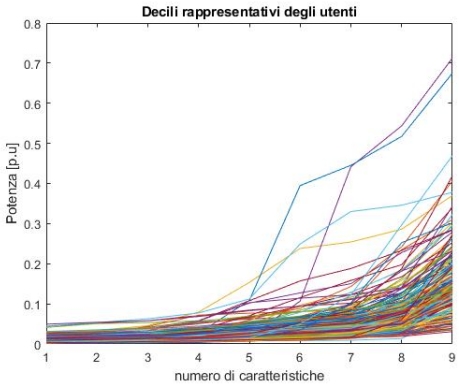

Figura 2: Decili rappresentativi degli utenti.

Applicazione a Dataset Reali

Descritti gli algoritmi e le fasi che precedono il clustering, si è passato all'applicazione su due dataset reali, in cui sono disponibili i consumi di utenti residenziali. Come prima cosa si è pensato di ripulire i dati, considerando solo le anomalie di tipo collettivo, confrontando i consumi giornalieri di ciascun mese. Non avendo un modello al quale fare riferimento sono stati eseguiti quattro metodi per la identificazione delle anomalie:

- Z-score
- MAD
- DBSCAN
- Analisi delle cumulative (test di Kolmogorov-Smirnov)

Ciascuno di essi ha riportato dei risultati buoni nel caso di anomalie estreme, mentre nel caso di utenti con consumi molto variabili nel mese solo l'analisi delle cumulative attraverso il test di Kolmogorov-Smirnov ha ottenuto risultati più solidi.

Infine si è applicata la metodologia di CONDUCT ai dataset ripuliti. Poiché entrambi i dataset non presentavano i valori di potenza contrattuale, si è deciso di cambiare e di utilizzare come potenza di riferimento il valore massimo presente all'interno del dataset.

La natura dei consumi degli utenti presenti in entrambi i dataset non ha permesso di risalire alla CDF delle differenze utile per ricavare i decili dalle curve di durata.

Molti degli utenti presentano dei consumi molto bassi con la presenza di pochi picchi molto pronunciati, in questo modo le differenze tra valori adiacente delle curve di durata risultano molto piccoli nella quasi totalità di punti, tranne per i valori finali dove si hanno differenza considerevoli per via appunto dei picchi.

Per questo motivo è stata utilizzata una CDF delle differenze diversa, che tiene conto delle variazioni di consumo durante la finestra temporale.

I decili ricavati dalla nuova CDF sono stati applicati come input degli algoritmi di clustering più classici:

- k-means.
- Clustering gerarchico, applicando diversi criteri di linkage.
- Spectral clustering (L), utilizza direttamente i decili come input.
- Spectral clustering (S), utilizza come input i valori di similarità, ricavati dai decili.

Indici di Validità del Clustering

I risultati ottenuti dagli algoritmi di clustering sono stati molto diversi tra di loro, ragione per cui sono stati utilizzati alcuni indici di validità del clustering ricavati dalla letteratura (Tabella I) per valutare sia la compattezza che la separazione tra cluster.

Tabella I: Indici di validità.

Algoritmi	**Thames valley dataset**				**Low Carbon London**			
	MIA	**CDI**	**SI**	**SMI2**	**MIA**	**CDI**	**SI**	**SMI2**
K-means	0.0147	0.141	2.15	0.879	0.0218	0.144	2.94	0.891
Hs	0.0276	0.262	4.21	0.822	0.0332	0.292	8.64	0.810
Hc	0.0178	0.145	2.43	0.841	0.0258	0.166	4.11	0.853
Ha	0.0188	0.166	2.37	0.827	0.0253	0.176	3.33	0.869
Hwa	**0.0136**	**0.127**	**2.03**	0.887	**0.0191**	**0.140**	**2.63**	0.881
Hwe	0.0182	0.156	2.31	0.833	0.0261	0.149	3.33	0.847
Spec-L	0.0296	0.670	3,85	**0.751**	0.0249	0.688	5,21	0.803
Spec-S	0.0357	0.353	7.40	0.815	0.0416	0.320	9.60	**0.792**

Una caratteristica comune degli indici di validità del clustering scelti è che valori inferiori corrispondono a soluzioni migliori. Dai valori degli indici riportati nella Tabella I risulta che la suddivisione ottenuta con il clustering gerarchico utilizzando il criterio di linkage di tipo Ward è quella che ha sfruttato al meglio la metodologia di CONDUCT. Il riconoscimento di utenti simili può essere utilizzato per lo studio di procedure di coinvolgimento degli utenti ai fini della variazione dei consumi, supportate da prezzi variabili o da incentivi diretti, come previsto nei programmi di *demand response*.